Comput. & Ops Res. Vol. 9, No. 2, pp. 119–126, 1982
Printed in Great Britain.

0305–0548/82/020119–08$03.00/0
Pergamon Press Ltd.

THE MEDIAN PROBLEM WITH CONGESTION

ODED BERMAN†
Faculty of Management, The University of Calgary, Calgary, Alberta,
Canada T2N 1N4

and

RICHARD C. LARSON‡
Operations Research Center, Massachusetts Institute of Technology, MA 02139, U.S.A.

Scope and Purpose—A central planning problem in both the public and private sectors is determining the best locations for a set of facilities. Each facility is assumed to house a number of servers who provide some service to a spatially dispersed set of customers. In many systems, customers are not scheduled—they arrive randomly in time—and the amount of service they require is also an uncertain quantity. In such cases, facility congestion can occur in the sense that all servers associated with one or more facilities become busy. While a considerable amount of effort has been devoted to optimal facility location, most of it has ignored this type of congestion, assuming essentially unlimited service capacity at each facility. In this paper we explicitly take facility congestion into account by allowing a customer to be serviced by servers from other than the closest facility—should the closest be saturated. In fact, depending on the pattern of facility congestion. any given customer may be serviced by a server from any one of the facilities. Given the model assumptions, the objective is to locate the facilities on the associated transportation network so as to minimize mean travel time to a random service request. This work begins to tie together previously disparate work in (deterministic) location theory and spatially distributed queues. In particular, the results are applicable to the hypercube queueing model, first reported seven years ago in this journal.

Abstract—The median problem has been generalized to include queueing-like congestion of facilities (which are assumed to have finite numbers of servers). In one statement of the generalizations, a closest *available* server is assumed to handle each service request. More general server assignment policies are admissable. The objective is to minimize the steady state expected travel time associated with a random service request. It is shown that, under suitable conditions, at least one set of optimal locations exists solely on the nodes of the network. It is also shown that this result has a direct relationship to the hypercube queueing model.

INTRODUCTION

The problem of determining the locations of a set of facilities on a network that minimize the expected travel time to or from the facilities, for the population of their users, is one of the classic problems in location theory. This problem, known in the literature as the *median problem*, has been studied very thoroughly in the last two decades. The basic theoretical results in this area are due to Hakimi[5, 6]. Subsequently, Goldman[4], Hakimi and Naheshwari[7], Levy[11] and Wendell and Hurter[13] have extended and generalized Hakimi's results.

When there are Q facilities to be located on an undirected network G, the median problem is to find a set $Z^* = (Z_1^*, Z_2^*, \dots, Z_Q^*)$ of Q points on G such that

$$\sum_{j=1}^{n} h_j d(Z^*, j) \leq \sum_{j=1}^{n} h_j d(Z, j), \tag{1}$$

†Oded Berman is an Associate Professor at the Faculty of Management, University of Calgary. He received his Ph.D. (1978) in Operations Research from M.I.T., and a B.A. in Economics and Statistics from Tel Aviv University. He taught at M.I.T. and was also a research staff member at the Electronic Systems Laboratory of M.I.T. Berman's papers were accepted for publication (or published) in the *European Journal of Operations Research, American Institute of Industrial Engineers Transactions, INFOR, Transportation Science, Computers and Operations Research, Networks, Decision Sciences*, and the *IEEE Transactions in Systems, Man, and Cybernetics.*

‡Dr. Larson is Prof. of Electrical Engineering and Computer Science and Urban Studies and Planning at the Massachusetts Institute of Technology. He is also Co-Director of the M.I.T. Operations Research Center and President of Public Systems Evaluation, Inc., a not-for-profit applied research organization. His most recent book is *Urban Operations Research* (co-authored with Amedeo R. Odoni), Prentice Hall, 1981. Dr. Larson's research interests involve the linkage of operations research methods with decision problems in the public sector. His papers have appeared in numerous journals, including *Operations Research, Management Science, Transportation Science, Networks, European Journal of Operational Research*, and *Journal of Criminal Justice.* Dr. Larson completed his contribution to this paper while a Visiting Professor at IMSOR (The Institute of Mathematical Statistics and Operations Research), The Technical University of Denmark, Lyngby, Denmark.

for all Q-tuples Z of points on G, where h_j is the fraction of demand that is generated at node j ($\Sigma_{j=1}^{n} h_j=1$), n is the number of nodes and $d(Z, j)$ is the shortest distance between node j and the closest point in Z. In [5] Hakimi proved that at least one such set Z^* exists solely on the nodes of the network, so that the search for an optimal solution can be confined to the finite node-set of G.

For this standard median problem to be applicable, four main assumptions should be satisfied:

(1) Travel in the given area is restricted to take place solely along the links of the network.

(2) Requests for service can occur only at the nodes of the network.

(3) When the number of facilities is greater than one, a service request from a particular location is always handled by a server at a closest facility.

(4) There is always an available (free) server at the selected (closest) facility.

The traveling associated with a service request could require the "customer" (requester of service) to travel to a nearest facility or a server at a nearest facility to travel to the customer. The former, "customer-to-server" type system, includes outpatient clinics, "little city halls", libraries and even hamburger havens. The latter, "server-to-customer" type system, includes emergency services (e.g. police, fire, ambulance, emergency repair), special-order delivery services, and certain home visitation medical services. In our work, we use the term "travel time associated with a service request" to mean either the customer-to-server or server-to-customer travel time.

The type of systems we consider are characterized by stochastically generated requests for service (in time and space) and by nondeterministic service times for meeting these requests. [A service time is comprised of travel time plus on-scene time.] In such an environment, it is quite possible that all servers at a nearest facility will be busy, thereby yielding a *congested network* in which queues could form. Thus, assumption (4) above often does not hold. For these systems, equation (1) expresses only an unnatural problem: finding a set of points so as to minimize the expected travel time for a random service request arising at one of those special times when servers are available at all the facilities. In contrast, it is the purpose of this work to incorporate in the median problem the possibility that all servers at any subset of the Q facilities can be busy.

The objective function in this *median problem with congestion* is to minimize the expected *response time* associated with a random service request, where response time is the sum of travel time and queueing delay. The averaging must be carried out using the equilibrium state probabilities of the system, defining "states" according to the status of each of the facilities—at least one server available at the facility or all servers busy. To avoid queue formation whenever possible, we assume that a *most preferred available server* is immediately dispatched to any service request. A request is placed in queue only if all servers are simultaneously busy; such a queue is assumed to be depleted in a first-in, first-out (FIFO) manner. Usually, server preferences are dependent solely on geographical proximity, but more general server preference policies are allowed.

In a recent paper[1] we solved this problem in its full generality for the case of one facility staffed with one server. Using results from the theory of $M/G/1$ queues, we developed an exact procedure for finding the optimal facility location, which was found to be sometimes on a node and oftentimes on a link. This result, which contrasts sharply with the nodal solution result of Hakimi and others, was due to nonlinearities in the mean queueing delay expression.

Our purpose in this paper is to extend the 1-server results to the case of multiple servers (and multiple facilities). However, due to the analytical intractability of $M/G/\eta$ queues (η = number of servers $\geq Q$), we have only been able to develop results for a special case, namely the case in which on-scene service time is much greater than travel time—so that state probabilities depend only on three sets of input parameters: nodal service request rates, *on-scene* service times, and server preferences. Changes of facility location occurring with all three parameter sets fixed do not change the state probabilities. However, this does not preclude dependence of system state probabilities on facility location: any change of facility location that creates a change in server preferences will, in general, change the state probabilities. With these restrictions, we prove a nodal result analogous to Hakimi's. However, it should be clear from the 1-server case that if the service time assumption is not valid in a particular setting, then optimal server locations are not in general at nodes.

The analysis, given the aforementioned restrictions, ties together previously disparate research efforts on network analysis and on spatial queueing analysis. In particular, we show that the hypercube model [9, 10] and the algorithm of Jarvis [8] for determining optimum locations can be useful to solve the median problem with congestion for specific situations. In addition this work indicates that the basic hypercube model does not suffer a loss of generality by considering only nodes (or atoms in the context of the hypercube model) for the locations of the service units.

NOTATIONS AND ASSUMPTIONS

Let $G(N, L)$ be an undirected network where N is the set of nodes with $|N| = n$, and L is the set of the links. Let $\bar{X}$ be the set of all possible locations of Q facilities ($Q > 1$), on the network G, i.e.

$$\bar{X}_Q = \{X_Q = (i_1, \ldots, i_Q); \qquad i_k \in G, \qquad k = 1, \ldots, Q\}.$$

Given any Q locations $X_Q = (i_1, \ldots, i_Q) \in \bar{X}_Q$, let $\hat{i}_k$ denote the event that the facility at i_k is not staffed with an available server (the facility is busy) and $\hat{\bar{i}}_k$ that the facility at i_k does have an available server. Therefore, for any $X_Q \in \bar{X}_Q$ there are 2^Q combinations (states) of finding the network at any time, according to the status of the Q facilities. Let $Y_{X(Q)}$ be the set of all states for $X_Q \in \bar{X}_Q$ and let $y_{X(Q)}$ (or for convenience y_Q) be a generic element of $Y_{X(Q)}$.

We assume that server assignment occurs according to a *fixed preference* procedure. That is, for each demand point in the network there is a list of facilities that specifies the ordering of preferences for the assignment of servers (i.e. first preference for servers from facility k_1, second preference for servers from facility k_2, etc.). A most preferred *available* server is always assigned to a customer.† The goal of the optimization to be stated below is to minimize expected system travel time under a given fixed preference procedure. The fixed preference itself need not be determined solely by relative travel times, but can include individual characteristics of servers (e.g. bilingualness) and special needs of customers at the nodes.

Let $t(i, j)$ be the travel time on link (i, j), $(i, j) \in L$, and let $d(y_Q, j)$ be the (minimum distance path) travel time associated with a most preferred available server to node j, when the system is in state y_Q.

As in the standard median problem we assume that service requests are generated on the nodes of the network. However, in addition, we assume that service requests occur according to a homogeneous Poisson process, with each request requiring a service time whose distribution is general but not dependent on its location or the location and identity of the server or the history of the system. This implies that on-scene service times are assumed to be *much greater* than travel times so that variations in total service times which are due solely to variations in travel times among potential servers are ignored. (This assumption is reasonable for systems whose on-scene service times are at least an order of magnitude greater than travel times.)

Finally, we require that travel time is uniform over a link, i.e. the travel time over a fraction θ of some link (p, q) is $\theta t(p, q)$. This assumption is not restrictive since the links and nodes can be defined in such a way that this assumption holds to a specified degree of accuracy.

MODEL FORMULATION AND ANALYSIS

We will consider the steady state behavior of the system. For any possible set of locations $X_Q \in \bar{X}_Q$, let $P(y_Q)$ be the steady state probability that the network is in state $y_Q \in Y_{X(Q)}$. (We assume that the appropriate ergodicity conditions apply so that a unique steady state distribution exists.). Let $y_Q{}^o$ be the state in which all the Q facilities are busy (i.e. $y_Q{}^o = (\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_Q)$ in our notation).

†When preferences depend directly on travel times, such a myopic strategy (through very reasonable) is not always optimal in the sense of minimizing time-average mean travel time. An optimal policy occasionally requires assignment of other than the most preferred available server [2], in order to leave the system in a state which best anticipates future service requests. We do not consider such strategies in our formulation of the median problem with congestion.

Conditioned on any state $y_Q \in Y_{X(Q)} - \{y_Q{}^o\}$, the expression

$$\sum_{j=1}^{n} h_j d(y_Q, j)$$

is the expected travel time associated with a random service request.

Suppose now that the network is in state $y_Q{}^o$. We will consider three alternative policies regarding this state:

(a) Service requests that occur while all the service units are busy, are handled by a back-up service system (zero line-capacity case). Let R be the travel time cost of utilizing this special reserve service system.

(b) Service requests that arrive while all the facilities are busy enter an infinite capacity queue that is depleted in a first-in, first-out (FIFO) manner; upon completion of service, the server is either assigned to the next request waiting in queue, or returns immediately home if none is waiting. Therefore, if $d(k, j) \equiv$ travel time between nodes k and j,

$$\sum_{k=1}^{n} \sum_{j=1}^{n} h_k h_j d(k, j)$$

is the expected travel time of a random service request given that the network is in state $y_Q{}^o$.

(c) ***Requires negative exponential service times.*** As in the previous case, service requests that arrive while all the facilities are busy enter a FIFO queue with infinite capacity, but now upon completion of service, the server always first returns to his/her home location. In this case, assuming negative exponential service times, the conditional expected travel time of a random request is

$$\sum_{k=1}^{Q} \sum_{j=1}^{n} \frac{\eta(k)}{\eta} h_j d(i_k, j),$$

given that the network is in state $y_Q{}^o$, where $\eta(k) =$ number of servers stationed at facility k, located at i_k, and $\eta = \Sigma_{k=1}^{Q} \hat{\eta}(k)$ is the total number of servers.

The appropriateness of any particular assumption depends, of course, on the system being modeled. Assumption (a) often applies to ambulance systems, in which emergency requests cannot be queued. Assumption (b) applies frequently to police vehicles that may be dispatched back-to-back to successive service requests. Assumption (c) applies to some ambulance and fire services, but suffers from the assumption of negative exponential service times. The median problem with congestion is now stated:

$$\min_{X_Q \in \bar{X}_Q} F(X_Q) \tag{2}$$

with

$$F(X_Q) = \sum_{y_Q \in Y_{X(Q)} - \{y_Q{}^o\}} P(y_Q) \sum_{j=1}^{n} h_j d(y_Q, j) + P(y_Q{}^o) \sum_{j=1}^{n} h_j C(j)$$

where

$$C(j) \text{ is } R \text{ or } \sum_{k=1}^{n} h_k d(k, j) \text{ or } \sum_{k=1}^{Q} \frac{\eta(k)}{\eta} d(i_k, j)$$

according respectively to (a), (b) or (c) above.

Obviously, the standard median problem is a special case of (2) arising when $P(y_Q) = 0$, $\forall\, y_Q \neq (\hat{i}, \ldots, \hat{i}_Q)$—the state where all the facilities have one or more units available *and* when $d(y_Q, j)$ is determined solely by geographic proximity (i.e. minimizing travel time). In contrast, the weights $P(y_Q)$ in (2) represent the fractions of time that the network is in each of the 2^Q

possible states. Therefore, as noted before, we take into account that any subset of facilities can become depleted of servers.

Now, given the assumptions stated before, the following theorem can be proved.

Theorem

For a *given* fixed preference server assignment procedure, at least one set of optimal solutions to (2) exists on the nodes of the network.

Proof. Let $X_Q = (i_1, i_2, \ldots, i_Q)$ be any feasible solution to (2), and let $P(y_Q)$, $\forall y_Q \in Y_{X(Q)}$ be the corresponding steady state probabilities. Suppose that i_S is an interior point on the link (p, q). Then by the uniform speed assumption

$$\frac{t(p, i_S)}{t(p, q)} = \theta \qquad 0 < \theta < 1. \tag{3}$$

The following proof is for the case $C(j) = \Sigma_{k=1}^{Q} (\eta(k)/\eta)\, d(i_k, j)$ in (2). The proofs for the other two cases are very similar and even slightly easier. Let $Y_{i_S} \subset Y_{X(Q)} - \{y_Q{}^o\}$ be the set of all states in which the facility located at i_S is available (i.e. has at least one available server). Then we can write $F(X_Q)$ as:

$$F(X_Q) = \sum_{y_Q \in Y_{i_S}} P(y_Q) \sum_{j=1}^{n} h_j d(y_Q, j) + P(y_Q{}^o) \left[\sum_{j=1}^{n} \frac{\eta(S)}{\eta} h_j d(i_S, j) \right] + A \tag{4}$$

where the term A includes all server assignments that exclude the facility located at i_S, i.e.

$$A = \sum_{y_Q \in Y_{X(Q)} - Y_{i_S} - \{y_Q{}^o\}} P(y_Q) \sum_{j=1}^{n} h_j d(y_Q, j) + P(y_Q{}^o) \left[\sum_{j=1}^{n} \sum_{\substack{k=1 \\ :k \neq S}}^{Q} \frac{\eta(k)}{\eta} h_j d(i_k, j) \right]. \tag{5}$$

Let $N_{y_Q}(i_S)$ be the set of all nodes that would be assigned to a server from the facility at i_S, when the network is in state $y_Q \in Y_{i_S}$, and let $\bar{N}_{y_Q}(i_S) = N - N_{y_Q}(i_S)$. Then we can rewrite (4) as:

$$F(X_Q) = \sum_{y_Q \in Y_{i_S}} P(y_Q) \left[\sum_{j \in N_{y_Q}(i_S)} h_j d(i_S, j) \right] + \frac{\eta(S)}{\eta} P(y_Q{}^o) \left[\sum_{j=1}^{n} h_j d(i_S, j) \right] + A + B \tag{6}$$

where the term B corresponds to non-queued assignment of servers not located at i_S, even when the facility located at i_S is available, i.e.

$$B = \sum_{y_Q \in Y_{i_S}} P(y_Q) \sum_{j \in \bar{N}_{y_Q}(i_S)} h_j d(y_Q, j). \tag{7}$$

Recalling that i_S is assumed to be an interior point on the link (p, q), let $N_{yQ}(i_S, p) \subset N_{yQ}(i_S)$ be the set of all nodes that belong to the set $N_{y_Q}(i_S)$ and which communicate most efficiently with the facility at i_S via p, and let $N_{yQ}(i_S, q) = N_{yQ}(i_S) - N_{yQ}(i_S, p)$. (The term "communicate" implies minimal travel time.) If a node communicates equally efficiently with i_S via nodes p or q for some y_Q, we can include that node in either $N_{y_Q}(i_S, p)$ or $N_{y_Q}(i_S, q)$, but not in both.

Let $N(i_S, p)$ be the set of *all* nodes which communicate most efficiently with the facility at i_S via node p and let $N(i_S, q) = N - N(i_S, p)$.

Therefore, we can write (6) as

$$\begin{aligned} F(X_Q) = \sum_{y_Q \in Y_{i_S}} P(y_Q) \Bigg\{ & \sum_{j \in N_{y_Q}(i_S, p)} h_j [d(j, p) + t(p, i_S)] \\ & + \sum_{j \in N_{y_Q}(i_S, q)} h_j [d(j, q) + t(q, i_S)] \Bigg\} + \frac{\eta(S)}{\eta} P(y_Q{}^o) \Bigg\{ \sum_{j \in N(i_S, p)} h_j [d(j, p) + t(p, i_S)] \\ & + \sum_{j \in N(i_S, q)} h_j [d(j, q) + t(q, i_S)] \Bigg\} + A + B. \end{aligned} \tag{8}$$

Using (3) and rearranging terms we get

$$F(X_Q)=\theta\Big[t(p,q)\Big(\sum_{y_Q\in Y_{i_S}}P(y_Q)\sum_{j\in N_{y_Q}(i_S,p)}h_j+\frac{\eta(S)}{\eta}P(y_Q{}^{o})\sum_{j\in N(i_S,p)}h_j\Big)\Big]$$
$$+(1-\theta)\Big[t(p,q)\Big(\sum_{y_Q\in Y_{i_S}}P(y_Q)\sum_{j\in N_{y_Q}(i_S,q)}h_j+\frac{\eta(S)}{\eta}P(y_Q{}^{o})\sum_{j\in N(i_S,q)}h_j\Big)\Big]$$
$$+A+B+C \tag{9}$$

where the term C corresponds to "fixed components" of travel time relating to the link (p, q), i.e.

$$C=\sum_{y_Q\in Y_{i_S}}P(y_Q)\Big[\sum_{j\in N_{y_Q}(i_S,p)}h_jd(j,p)+\sum_{j\in N_{y_Q}(i_S,q)}h_jd(j,q)\Big]$$
$$+\frac{\eta(S)}{\eta}P(y_Q{}^{o})\Big[\sum_{j\in n(i_S,p)}h_jd(j,p)+\sum_{j\in N(i_S,q)}h_jd(j,q)\Big]. \tag{10}$$

Since the steady state probabilities are independent of θ (refer to the assumption from previous section), $F(X_Q)$ is a concave function of θ. This implies that the minimum of $F(X_Q)$ as a function of θ occurs at an extreme point, either $\theta=0$ or 1, corresponding to locations at node p or q, respectively. Clearly, the node p is optimal if the coefficient of θ in (9) is larger than the coefficient of $(1-\theta)$; otherwise q is optimal or a tie exists, in which case either is optimal. We showed so far that by starting with any X_Q which is not at all nodes, there is an $X_Q{}^1$ with $F(X_Q{}^1)\le F(X_Q)$ which agrees with X_Q except that one non-node point i_S of X_Q has been shifted to one of the two nodes adjacent to it. This argument can be repeated to show that there is an all-node X_Q^2 with $F(X_Q{}^2)\le F(X_Q)$. Thus, the search for an optimum can be restricted to nodes; since there are only finitely many sets of Q nodes, the minimum must be attained (minimum instead of merely infimum).

THE MEDIAN PROBLEM WITH CONGESTION AND THE HYPERCUBE MODEL

"The hypercube model" is a spatially distributed queueing model developed by Larson[9] to investigate analytically the performance of urban emergency services. The model assumes a geographical region R that is divided into n geographic areas or atoms. The fraction of demand associated with each atom j is $h_j(\Sigma_{j=1}^{n} h_j = 1)$ and the travel time from atom i to atom j is $d(i, j)$. Service requests over the entire region are generated in a Poisson manner at a rate λ and at each atom j independently in a Poisson manner with rate $\lambda_j(\Sigma_j \lambda_j = \lambda)$.

There are a total of Q service units (i.e. servers) to respond to the requests for service, located at atoms $i_1, i_2, \ldots, i_Q$. For Markov analysis, the service time for each unit n is assumed to be exponential with mean $\mu_n{}^{-1}$. Recent research has shown that the assumption of exponentiality of the service time does not markedly affect the predictive accuracy of the model[8].

States of the system are defined by the status of each service unit: busy or available. The model allows a zero line-capacity (implying the existence of special reserve units), or an infinite capacity queue. Given a dispatching policy, all the 2^Q steady state probabilities of the system can be obtained by solving 2^Q detailed balance equations[9]. In [10], Larson used a server sampling scheme adapted from the $M/M/Q$ model to obtain fast approximate solutions for the required probabilities.

For a given set of single server locations at atoms $i_1, \ldots, i_Q$ the hypercube model computes several performances measures. Among them, the most imporant one is the region-wide mean travel time, defined as

$$\sum_{j=1}^{n}\sum_{k=1}^{Q}\rho_{i_k,j}\,d(i_k,j)+P\text{ (all units are busy)}\sum_{j=1}^{n}h_j\,C(j) \tag{11}$$

where $\rho_{ik,j}\equiv$fraction of all nonqueued dispatches that send the unit from atom i_k to atom j;

$k=1,\ldots,Q$; $j=1,\ldots,n$; $C(j)\equiv$average travel time to node j arising from dispatches from queued service requests (infinite capacity case) or from service requests handled by a back-up service system. The $\rho_{ik,j}$'s represent the ***response patterns*** of units.

In [8] Jarvis developed an algorithim to find a set of "optimum" locations in the framework of the hypercube model where locations are constrained to atoms. The key idea behind the Jarvis algorithm is to optimally locate the servers (facilities) for a given response pattern and then, given a new set of locations, to reassess the response patterns to determine if a new set of dispatch preferences (and thus response patterns) could improve system performance further. This alternating iterative procedure is analogous to the "locate–allocate" scheme often used in deterministic location theory[12].

Jarvis' algorithm for the zero line-capacity case works as follows:

(1) *Initialize*: Specify initial unit locations for units $1,2,\ldots,Q$, corresponding to atoms $i_1, i_2,\ldots, i_Q$.

(2) *Allocate*: Solve the hypercube model to obtain $\rho_{i_k,j}$, $k=1,\ldots,Q$; $j=1,\ldots,n$.

(3) *Locate*: Solve the following L.P problem:

$$\min \sum_{k=1}^{Q}\sum_{v=1}^{n} P(v,k)\,C(v,k)$$

$$\text{s.t.} \sum_{v=1}^{n} P(v,k)=1 \qquad k=1,\ldots,Q$$

$$P(v,k)\ge 0 \qquad v=1,\ldots n;\ k=1,\ldots,Q$$

where the decision variable $P(v,k)$ is the probability that server k is at node v when available, $v=1,\ldots,n$; $k=1,\ldots,Q$; and $C(v,k)=\Sigma_{j=1}^{n}\rho_{i_k,j}\,d(v,j)$, $v=1,\ldots,n$; $k=1,\ldots,Q$. The LP has an optimal solution with $P(v,k)=1$ for some v and every k.

(4) *Test for convergence*: If the new Q locations are identical to the old set of Q locations, *stop*. Otherwise, go to Step 2 with new unit locations $i_1,\ldots,i_Q$, and reallocate.

Whenever the algorithm terminates, at least a local optimal solution is ensured. By taking several different initial sets of locations, the chances of getting closer to the optimal global solution are improved.

The hypercube model can be applied in our congested median network context. The network G can represent the geographical region R, the nodes of the network being the atoms, and the links being the major streets connecting the atoms. It requires only some algebraic manipulations to demonstrate that if we take any Q points in the network to be the set of server locations, then $F(X_Q)$—the cost function of our problem (2)—turns out to be identical to the mean region-wide travel time of the hypercube model[9].

The conclusion of this discussion is that since the assumptions of our Theorem hold for the hypercube model (subject to our discussion of service times), neither the hypercube model nor Jarvis' algorithm suffer a loss of generality by considering only locations on the atoms. In addition Jarvis' algorithm can be applied to the median problem with congestion whenever the hypercube model's assumptions are accepted. This result ties together two very diпerent approaches in location theory, one which is purely deterministic (the median problem) and the other stochastic (the hypercube model).

EXAMPLE

The following example will illustrate some of our previous discussion. Suppose we want to locate three facilities on the simple network show‥ in Fig. 1. The numbers next to the nodes are the fractions of demands from each node h_j; $j=1,\ldots,5$ and the numbers next to the links are the travel times. There are $\binom{5}{3}$ possible distinct locations:

$$\{1,2,3\},\quad \{1,2,4\},\quad \{1,2,5\},\quad \{1,3,4\},\quad \{1,3,5\},\quad \{1,4,5\}$$

$$\{2,3,4\},\quad \{2,3,5\},\quad \{2,4,5\},\quad \{3,4,5\}.$$

The optimal location according to the standard 3-median problem is $\{1,2,5\}$, which can be

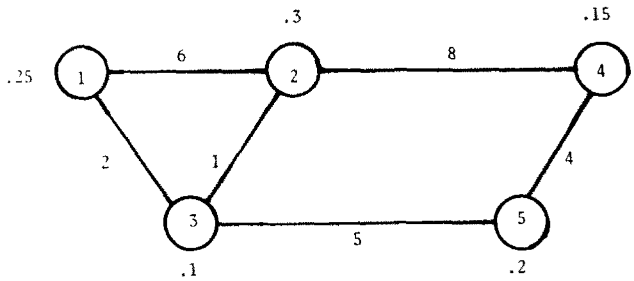

Fig. 1. A simple 5-node network.

obtained by hand. Suppose, however, that service requests occur in the network in a Poisson fashion with $\lambda = 4$, and the service time for each one of the three units is exponential with identical means $\mu^{-1} = 1$. Let us assume that $R = 5$ units of time—the cost resulting when dispatching a reserve unit. We also assume that server preferences are determined solely by geographical proximity.

The Jarvis algorithm with an initial location at the absolute 3-median, i.e. $\{1,2,5\}$, converges after one iteration to the optimal solution at location $\{2,3,5\}$. The improvement achieved by moving from the location $\{1,2,5\}$ to $\{2,3,5\}$ is 3% in terms of the median problem with congestion. It is interesting to observe that the location $\{2,3,5\}$ is among the weakest possible locations in terms of the standard median problem (only location $\{1, 2, 3\}$ is worse). This indicates that blind application of the absolute (deterministic) median problem to stochastic situations can lead to erroneous results, even for such simple networks.

REFERENCES

1. O. Berman and R. C. Larson, Optimal server location on a network operating as an *M/G/*1 queue. *Ops Res.* (Submitted).
2. G. M. Carter, J. M. Chaiken and E. Ignall, Response areas for two emergency units. *Ops Res.* **40**(3), 571–574 (1972).
3. L. Cooper, Solutions of generalized locational-equilibrium problems. *J. Reg. Sci.* **7**, 1–18 (1969).
4. A. J. Goldman, Optimum locations for centers in networks. *Transportation Sci.* **3**(4), 352–360 (1969).
5. S. L. Hakimi, Optimal locations of switching centers and absolute centres and medians of a graph. *Ops Res.* **12**, 450–459 (1964).
6. S. L. Hakimi, Optimum distribution of switching centers on a communications network and some related graph theoretic problems. *Ops Res.* **13**, 462–475 (1965).
7. S. L. Hakimi and S. H. Hasheshwari, Optimum locations on centers in networks. *Ops Res.* **20**, 967–977 (1977).
8. J. P. Jarvis, Optimization in stochastic service with distinguishable servers. *TR*-19-75, Innovative Resource Planning Project in Urban Public Safety Systems, Operations Research center, M.I.T. (June 1975). Also J. P. Jarvis, A. location model for spatially distributed queueing systems. *Proc. Int. Conf. on Cybernetics and Society*, pp. 32–35, November 1976.
9. R. C. Larson, A hypercube model for facility location and redistricting in urban emergency services. *Comput. Ops Res.* **1**, 67–95 (1974).
10. R. C. Larson, Approximating the performance of urban emergency service systems. *Ops Res.* **23**(5), 845–868 (1975).
11. J. Levy, An extended theorem for location on a network. *Ops Res. Quart.* **18**(4), 433–442 (1967).
12. Y. Rapp, Planning of exchange locations and boundaries in multi-exchange networks. *Ericsson Technics* **18**(2), 91–113 (1962).
13. R. E. Wendell and A. P. Hurter, Optimal location on a network. *Transportation Sci.* **7**(1), 18–33 (1973).